Statistica Sinica **15**(2005), 635-644

LIKELIHOOD RATIO TESTS UNDER LOCAL ALTERNATIVES IN REGULAR SEMIPARAMETRIC MODELS

Moulinath Banerjee

University of Michigan

Abstract: We consider the behavior of likelihood ratio statistics for testing a finite dimensional parameter, or functional of interest, under local alternative hypotheses in regular semiparametric problems. These are problems where $\sqrt{n}$–regular estimates of the parameter/functional of interest exist and, in particular, the MLE converges at $\sqrt{n}$ rate to the true value and is asymptotically normal and efficient. We show that in regular problems, the likelihood ratio statistic for testing $H_0 : \theta(\psi) = \theta(\psi_0) = \theta_0$ (where ψ_0 is a fixed point in the infinite–dimensional parameter space Ψ and $\theta(\psi)$ is a finite-dimensional (sub)parameter or functional of interest) converges in distribution under local (contiguous) alternatives of the form $\psi_n = \psi_0 + n^{-1/2}\,h + o(n^{-1/2})$ to a non-central χ^2 random variable, with non-centrality parameter involving the direction of perturbation h and the efficient information matrix for θ under parameter value ψ_0. This conforms to what happens in the case of regular parametric models in classical statistics.

Key words and phrases: Asymptotic distribution, χ^2 distribution, confidence sets, contiguity, Cox model, least favorable submodels, likelihood ratio, local alternatives.

1. Introduction

Let $X_1, \ldots, X_n$ be a random sample from the distribution P_ψ, ψ belonging to a set Ψ. We assume a topology on the set Ψ; in applications Ψ is usually a subset of a Banach or Hilbert space. Let $\mathrm{lik}(\psi, x)$ denote the likelihood for an observation. In contrast to classical parametric statistics, $\mathrm{lik}(\psi, x)$ in many cases is a modification of $p(\psi, x)$ where $p(\psi, \cdot) = dP_\psi/d\mu$, μ being some common dominating σ - finite measure. The likelihood ratio statistic (henceforth LRS) for testing the null hypothesis $H_0 : \theta(\psi) = \theta_0$ for some functional θ is given by:

$$\begin{aligned}\mathrm{lrt}_n(\theta_0) &= 2\Big\{\sup_{\psi\in\Psi}\Big(\sum_{i=1}^{n}\log\ \mathrm{lik}(\psi, X_i)\Big) - \sup_{\psi:\theta(\psi)=\theta_0}\Big(\sum_{i=1}^{n}\log\ \mathrm{lik}(\psi, X_i)\Big)\Big\}\\ &= 2n\mathbb{P}_n(\log(\mathrm{lik}(\hat{\psi}))) - 2n\mathbb{P}_n(\log(\mathrm{lik}(\hat{\psi}_0)))\,;\end{aligned}$$

here $\hat{\psi}$ is the unconstrained maximizer of the log–likelihood and $\hat{\psi}_0$ is the maximizer under the constraint imposed by H_0, while $\mathbb{P}_n$ denotes the empirical measure of the observations $X_1, \ldots, X_n$. In the cases we consider, θ takes values in $\mathbb{R}^k$.

The behavior of $\mathrm{lrt}_n(\theta_0)$ under H_0, for a 1–dimensional parameter θ, was established in Theorem 3.1 of Murphy and Van der Vaart (1997) for a broad class of semiparametric models characterized by the requirement that the MLE of θ converges at the "regular" $\sqrt{n}$ rate to a normal limit and admits an asymptotic linear representation. It was shown that the LRS converges to a χ_1^2 distribution. This approach extends readily to the case of a vector–valued parameter (here, the limit distribution is χ_k^2, k being the dimension of θ) and was investigated by Murphy and Van der Vaart (2000) in a profile–likelihood setting. In this paper we focus on the behavior of the LRS in these regular semiparametric models under a sequence of local alternatives that converge to a point in the null hypothesis with increasing sample size. More specifically, let $\psi_n = \psi_0 + (1/\sqrt{n})h + o(1/\sqrt{n})$, where $\theta(\psi_0) = \theta_0$. We study the limiting behavior of $\mathrm{lrt}_n(\theta_0)$ based on i.i.d. data $X_1, \ldots, X_n$, when P_{ψ_n} is taken to be the true underlying distribution at stage n. The alternatives $\{\psi_n\}$ correspond to the curve $\psi(t) \equiv \psi_0 + t\,h + o(t)$ with gradient h at the point 0. We show that the likelihood ratio statistic is asymptotically distributed as a non–central χ_k^2 random variable with non–centrality parameter Δ which involves I_0, the efficient information for the finite–dimensional parameter θ under parameter value ψ_0, and the direction of perturbation of ψ_0. In the case where the parameter ψ can be partitioned as (θ, η) (with η being infinite–dimensional), the expression for the non–centrality parameter matches that obtained for regular parametric models.

Before proceeding to the key results in the next section, we impose some structural requirements on the underlying model $\{P_\psi : \psi \in \Psi\}$ and the functional of interest θ.

Hellinger Differentiability: Assume that the parameter ψ lies in a subset Ψ of some Hilbert space $\mathcal{H}$. Consider the path in Ψ given by $\psi_t = \psi + h\,t + o(t)$; thus h is the gradient of the path at $t = 0$. Let $s(\psi, \cdot) = (dP_\psi/d\mu)^{1/2}$, where μ is a σ-finite measure dominating the P_ψ's. The Hellinger differentiability condition can then be stated as:

$$\int \left[\frac{s(\psi_t, \cdot) - s(\psi, \cdot)}{t} - \frac{1}{2}(Ah)s(\psi, \cdot)\right]^2 d\mu(\cdot) \to 0 \ \text{ as } \ t \to 0\,. \tag{1.1}$$

Here A is a bounded linear operator defined on $\dot{\mathcal{H}}$, the closed linear span of the h's (the gradients of the ψ_t's) and assumes values in $L_2^0(P_\psi)$. In fact, Hellinger differentiability implies that the range of A is contained in $L_2^0(P_\psi)$, the subset of square integrable functions under the measure P_ψ that have mean 0 (see, for example, Section 2.1 of Bickel, Klaassen, Ritov and Wellner (1998)).

Differentiability of θ: The functional of interest θ, treated as a function of ψ, i.e., $\theta : \Psi \to \mathbb{R}^k$, is differentiable in the sense that $(\partial/\partial t)(\theta(\psi_t)) \mid_{t=0} = L(h)$ where L is a bounded linear map from $\dot{\mathcal{H}} \to \mathbb{R}^k$. Observe that by the characterization of real-valued functionals defined on Hilbert spaces, we have $L(h) = (\langle \dot\theta_{10}, h\rangle, \ldots, \langle \dot\theta_{k0}, h\rangle)^T$ for some $\dot\theta_{10}, \ldots, \dot\theta_{k0}$ all belonging to $\dot{\mathcal{H}}$.

2. The Asymptotic Distribution of the Likelihood Ratio Statistic: Main Results

We first formulate an extension of Theorem 3.1 of Murphy and Van der Vaart (1997) (dealing with the behavior of $\mathrm{lrt}_n(\theta_0)$ under the null hypothesis) that applies to vector–valued functionals θ of any dimension. This is exploited along with results from contiguity theory to derive the limit distribution of the likelihood ratio statistic under a sequence of local alternatives of the type considered in the previous section. The use of contiguity theory stems from the fact that the model P_ψ is Hellinger differentiable along the curve P_{ψ_t} which implies, through a LAN (local asymptotic normality) expansion of the log-likelihood ratio log $d\,P^n_{\psi_n}/d\,P^n_{\psi_0}$, that the sequence of probability measures $\{P^n_{\psi_n}\}$ (the n–fold product of P_{ψ_n}) and $\{P^n_{\psi_0}\}$ (the n-fold product of P_{ψ_0}) are mutually contiguous.

The General Semiparametric Likelihood Ratio Statistic Theorem: Suppose that ψ_0 is the true value of the parameter with $\theta(\psi_0) = \theta_0$ (thus $\Psi_0 \in H_0$) and denote P_{ψ_0} by P_0.

A.1 Assume that as $n \to \infty$ the MLE $\hat\theta = \theta(\hat\psi)$ satisfies

$$\sqrt{n}(\hat\theta - \theta_0) = \sqrt{n} I_0^{-1} \mathbb{P}_n(\tilde l) + o_p(1) \tag{2.1}$$

under P_0, where $P_0(\tilde l) = 0$ and $I_0 = P_0(\tilde l\, \tilde l^T)$. Thus $\tilde l \in L^0_2(P_0)^k$.

A.2 Assume that, for every t in a neighborhood U of θ_0 and every ψ in a neighborhood V of ψ_0, there exists a surface $t \mapsto \xi(t,\psi)$ taking values in Ψ that satisfies the following:

(a) $\theta(\xi(t,\psi)) = t$;

(b) $\xi(t,\psi) \mid_{t=\theta(\psi)} = \psi$;

(c) $t \mapsto l(x;t,\psi) = \ln(\mathrm{lik}(\xi(t,\psi),x))$ is twice continuously differentiable in t for every x with derivatives $\dot l$ and $\ddot l$ with respect to t;

(d) $\dot l(\cdot;\theta_0,\psi_0) = \tilde l$ and $-\mathbb{P}_n(\ddot l(\cdot;\tilde\theta,\tilde\psi)) \to_P P_0(\tilde l\, \tilde l^T) = I_0$ for any random $\tilde\theta \to_P \theta_0$ and $\tilde\psi \to_P \psi_0$ under P_0 and $\sqrt{n}\, \mathbb{P}_n(\dot l(\cdot;\theta_0,\hat\psi_0) - \tilde l) \to_{P_0} 0$.

A.3 Suppose that both the unconstrained and constrained maximizers of the likelihood, $\hat\psi$ and $\hat\psi_0$, are consistent under P_0.

Theorem 2.1. *If* A.1, A.2 *and* A.3 *hold, then*

$$\mathrm{lrt}_n(\theta_0) = \sqrt{n}\left(\hat{\theta}_n - \theta_0\right)^T I_0 \sqrt{n}\left(\hat{\theta}_n - \theta_0\right) + o_p(1) \to_d Z^T I_0 Z \sim \chi^2_k,$$

where $Z \sim N_k(0, I_0^{-1})$.

The proof of this theorem is skipped. The function $\tilde{l}$ that appears in (2.1) is actually the efficient score function in regular semiparametric models. For the details, see Banerjee (2000, Chap.2).

Power under Local (Contiguous) Alternatives. The following theorem characterizes the limiting power of the likelihood ratio test under a sequence of local alternatives converging to the true value of the parameter.

Theorem 2.2. *Consider the likelihood ratio statistic* $lrt_n(\theta_0)$ *for testing the null hypothesis* $\theta = \theta_0 \equiv \theta(\psi_0)$. *Assume that the model is Hellinger differentiable along the curve* $\psi_t = \psi_0 + t\,h + o(t)$ *with derivative operator* A. *Also assume that conditions* A.1 *to* A.3 *of Theorem* 2.1 *hold under* P_{ψ_0}. *Then, under the sequence of local alternatives* $P_{\psi_0 + h/\sqrt{n} + o(1/\sqrt{n})}$, *the likelihood ratio statistic* $lrt_n(\theta_0)$ *converges in distribution to* L *where* L *follows a* χ^2_k *distribution with non-centrality parameter* $c^T I_0\, c$; *here,* I_0 *is the covariance matrix of* $\tilde{l}$ *and* c *is the covariance between* $A\,h$ *and* $\tilde{l}$ *in* $L_2^0(P_{\psi_0})$ *scaled by* I_0, *or* $c = I_0^{-1}\langle Ah\,,\, \tilde{l}\rangle_{P_{\psi_0}}$.

Comments: Note that A in (1.1) is taken to be a bounded linear (score) operator defined on the closed linear span of gradients. However, the above theorem goes through without the requirement of a score operator so long as we have Hellinger differentiability along the path $\psi_0 + h\,t + o(t)$ with some $g \in L_2^0(P_{\psi_0})$ playing the role of $A\,h$ in (1.1). The conclusion in the statement of Theorem 2.2 remains valid with $A\,h$ replaced by g.

Proof of 2.2. Hellinger differentiability of the model along the curve ψ_t implies that

$$\int \Big[\frac{dP_{\psi_t}^{1/2} - dP_{\psi_0}^{1/2}}{t} - \frac{1}{2}(Ah)dP_{\psi_0}^{1/2}\Big]^2 \to 0$$

as $t \to 0$. Now set

$$\Lambda_n = \log\Big(\frac{dP^n_{\psi_0 + \frac{h}{\sqrt{n}} + o(\frac{1}{\sqrt{n}})}}{dP^n_{\psi_0}}\Big).$$

This is the log likelihood ratio based on observations $X_1, \ldots, X_n$ at stage n . By Lemma 3.10.11 of Van der Vaart and Wellner (1996), we get a LAN expansion for the log–likelihood ratio as

$$\Lambda_n = \log\frac{dP^n_{\psi_0 + \frac{h}{\sqrt{n}} + o(\frac{1}{\sqrt{n}})}}{dP^n_{\psi_0}} = \frac{1}{\sqrt{n}}\sum_{i=1}^n Ah(X_i) - \frac{1}{2}\|Ah\|^2 + o_{P^n_{\psi_0}}(1)\,.$$

It follows that $\Lambda_n \to_d W$ where $W \sim N(-(1/2)\|Ah\|^2, \|Ah\|^2)$, under $P^n_{\psi_0}$. Therefore, under $\{P_{\psi_0^n}\}$,

$$\exp(\Lambda_n) \equiv \frac{dP^n_{\psi_0+\frac{h}{\sqrt{n}}+o(\frac{1}{\sqrt{n}})}}{dP^n_{\psi_0}} \to_d \exp W\,.$$

But $E(\exp(W)) = \exp(-(1/2)\,\sigma^2 + (1/2)\,\sigma^2) = 1$, using the formula for the moment generating function of the normal distribution. By Le Cam's first lemma (Van der Vaart (1998, p.88)), we conclude that the sequences of probability measures $\{P^n_{\psi_0+h/\sqrt{n}+o(1/\sqrt{n})}\}$ and $\{P^n_{\psi_0}\}$ are contiguous. Consequently the convergences in probability that hold under $P^n_{\psi_0}$ also hold under $P^n_{\psi_0+h/\sqrt{n}+o(1/\sqrt{n})}$, and vice-versa.

Consider now the joint distribution of $T_n = \sqrt{n}\,\mathbb{P}_n I_0^{-1}\tilde{l}$ and Λ_n under $P^n_{\psi_0}$. From

$$\begin{pmatrix} T_n \\ \Lambda_n \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{n}}\sum I_0^{-1}\tilde{l}(X_i) \\ \frac{1}{\sqrt{n}}\sum Ah(X_i) - \frac{1}{2}\|Ah\|^2 \end{pmatrix} + o_{P^n_{\psi_0}}(1)\,,$$

the CLT in conjunction with Slutsky's theorem yields:

$$\begin{pmatrix} T_n \\ \Lambda_n \end{pmatrix} \to_d N_{k+1}(\mu, \Sigma)$$

under $P^n_{\psi_0}$ with $\mu = (0_{k\times 1}, -(1/2)\|Ah\|^2)^T$, and

$$\Sigma = \begin{pmatrix} I_0^{-1} & c^T \\ c & \|Ah\|^2 \end{pmatrix}.$$

Here $c = \mathrm{Cov}\,(Ah, I_0^{-1}\tilde{l}) = I_0^{-1}\,\mathrm{Cov}\,(Ah, \tilde{l}) = I_0^{-1}\langle Ah\,,\,\tilde{l}\rangle_{P_{\psi_0}}$. Note that $\langle Ah, \tilde{l}\rangle_{P_{\psi_0}} = (\int (Ah)\tilde{l}_1 dP_{\psi_0}, \ldots, \int (Ah)\tilde{l}_k dP_{\psi_0})^T$. Now by Le Cam's third lemma (Van der Vaart (1998, p.90)), the limiting distribution of $(T_n, \Lambda_n)^T$ under the sequence $P^n_{\psi_0+h/\sqrt{n}+o(1/\sqrt{n})}$ is $N_{k+1}(\tilde{\mu}, \Sigma)$, where $\tilde{\mu} = (c, (1/2)\|Ah\|^2)^T$. This shows that $T_n \to_d N(c\,,\,I_0^{-1})$ under $P^n_{\psi_0+(h/\sqrt{n})+o(1/\sqrt{n})}$. We now note that conditions A.2 and A.3 in Theorem 2.1 continue to hold under $P^n_{\psi_0+h/\sqrt{n}+o(1/\sqrt{n})}$, and condition A.1 is modified to $\sqrt{n}(\hat{\theta} - \theta_0) = T_n + o_p(1)$ under $P^n_{\psi_0+(h/\sqrt{n})+o(1/\sqrt{n})}$ with T_n converging in distribution to $N(c,\,I_0^{-1})$. Consequently $\mathrm{lrt}_n(\theta_0)$, which in this case is still $\sqrt{n}(\hat{\theta} - \theta_0)^T I_0 \sqrt{n}(\hat{\theta} - \theta_0) + o_p(1)$, converges to $\chi^2_k\,(c^T I_0 c)$.

The covariance vector c is the derivative of the curve $\theta(\psi_t)$ at $t = 0$ when θ is "pathwise norm differentiable" in the sense of Van der Vaart (1991). To see this, we briefly review the concept of pathwise norm differentiability. Consider θ as a functional from the space of probabilities to $\mathbb{R}^k$; hence write $\theta(\psi_t) \equiv \theta(P_{\psi_t})$.

Then θ is called *pathwise norm-differentiable* if there exists a bounded linear map $\dot{\theta}$ from $\overline{\mathcal{R}(A)} \to \mathbb{R}^k$ such that

$$L(h) = \lim_{t\to 0}\left[\frac{\theta(P_{\psi_t}) - \theta(P_\psi)}{t}\right] = \dot{\theta}\,(A\,h)\ .$$

A necessary and sufficient condition for pathwise norm differentiability due to Van der Vaart (see Van der Vaart (1991)) is given by $\mathcal{R}(L^*) \subset \mathcal{R}(A^*)$, and failure of this condition implies the non-existence of $\sqrt{n}$–regular estimators of θ. When the condition for differentiability is satisfied we have that $L = \dot{\theta} \circ A$ so that $L^* = A^* \circ \dot{\theta}^*$. It is easily shown that the above necessary and sufficient condition is equivalent to the following: for each $1 \leq j \leq k$ there exists a (necessarily unique) solution to the equation $A^* x = \dot{\theta}_{j0}$ in $\overline{\mathcal{R}(A)}$. We denote the unique vector of solutions by $g_0 = (\dot{g}_{10}, \ldots, \dot{g}_{k0})$. Since it is also the case that for each j, $\dot{\theta}_{j0} = L^*(e_j) = A^*(\dot{\theta}^*(e_j))$ (where e_j denotes the j'th canonical basis vector) and $\dot{\theta}^*$ assumes values in $\overline{\mathcal{R}(A)}$, we conclude that $\dot{g}_{j0} = \dot{\theta}^*(e_j)$. We call $\dot{\theta}^*$ the efficient influence function and identify it with the vector g_0, its values at the canonical basis vectors. In many semiparametric situations it is this vector that provides the linear approximation to the centered and scaled MLE of θ, the parameter of interest, i.e., $g_0 = I_0^{-1}\,\tilde{l}$.

Corollary 2.1. *If θ is pathwise norm-differentiable and the efficient influence function for the estimation of θ at ψ_0 , $g_0 = I_0^{-1}\,\tilde{l}$, then $c = (\partial/\partial t)\theta(\psi_t)\mid_{t=0}$.*

Proof. We have $c = \mathrm{Cov}\,(Ah\,,\, I_0^{-1}\tilde{l}) = \mathrm{Cov}\,((\dot{g}_{10}, \ldots, \dot{g}_{k0})^T\,,\, Ah) = (\langle \dot{g}_{10}\,,\, Ah\rangle, \ldots, \langle \dot{g}_{k0}\,,\, Ah\rangle)^T = (\langle A^*\dot{g}_{10}\,,\, h\rangle, \ldots, \langle A^*\dot{g}_{k0}\,,\, h\rangle)^T = (\langle \dot{\theta}_{10}\,,\, h\rangle, \ldots, \langle \dot{\theta}_{k0}\,,\, h\rangle)^T = \frac{\partial}{\partial t}\theta(\psi_t)\mid_{t=0}$.

For yet another characterization of c in terms of least favorable directions, see Banerjee (2000, Chap.2). Also note that Theorem 2.2 and Corollary 2.1 remain valid for a Banach space valued parameter ψ (pathwise norm differentiability is characterized in exactly the same way as in the Hilbert space case if adjoints are given the proper interpretations).

3. Partitioned Parameters

Here we use the results of the previous section to obtain expressions for the power of the likelihood ratio test under local alternatives when the parameter is naturally partitioned into two components: the first, θ, which is also the parameter of interest, belonging to a Euclidean space and the second, η, being infinite dimensional. This situation *arises extensively* in semiparametric models (and is therefore deemed important in its own right), with perhaps the most common example being the Cox Proportional Hazards setting where the hazard function corresponding to the survival time of an individual is modeled as

$\Lambda(t \mid Z) = e^{\theta^T Z}\Lambda(t)$. Here the parameter vector is $\psi = (\theta, \Lambda)$, with θ the Euclidean regression parameter measuring the effect of measured covariates on the survival time, and Λ the baseline hazard function taking values in the space of all cadlag functions defined on a bounded interval. Other useful models, where the parameter is naturally partitioned, include frailty models (for example, the Gamma frailty models considered in Murphy and Van der Vaart (1997), Murphy and Van der Vaart (2000) and the exponential frailty model also considered in these papers), case control studies for missing covariates (Roeder, Carroll and Lindsay (1996), Murphy and Van der Vaart (2000) and Murphy and Van der Vaart (2001)), partially linear regression models (see, for example Green (1984), Engle et al. (1986) and Van der Vaart (1998)) and semiparametric mixture models.

Let $\mathcal{P} = \{P_{\theta,\eta} : \theta \in \Theta, \eta \in \mathcal{H}\}$, where Θ is an open subset of $\mathbb{R}^k$ and $\mathcal{H}$ is some subset of a Banach space $\mathcal{G}$. Consider a fixed set of paths in $\mathcal{H}$ of the form $\eta_t = \eta + \beta t + o(t)$ with $\beta \in \mathcal{G}$. Let $\mathcal{C}$ denote the closed linear span of the β's. Now consider paths of the form $(\theta + th, \eta_t) \in \Theta \times \mathcal{H}$, and assume Hellinger differentiability with respect to this set of paths. Thus

$$\lim_{t\to 0}\int \Big[\frac{dP^{1/2}_{\theta+th,\eta_t} - dP^{1/2}_{\theta,\eta}}{t} - \frac{1}{2}(\dot{l}_\theta^T h + \dot{l}_\eta\beta)dP^{1/2}_{\theta,\eta}\Big]^2 = 0\,.$$

Now, $\mathbb{R}^k \times \mathcal{C}$ is a Banach space itself with the product topology and in this situation the score operator $A : \mathbb{R}^k\times\mathcal{C} \to L_2^0(P_{\theta,\eta})$ is given by $A\,(h,\beta) = \dot{l}_\theta^T h + \dot{l}_\eta\beta$, where $\dot{l}_\theta \in L_2^0(P_{\theta,\eta})^k$ is the score function for θ and the score operator for η, the nuisance parameter, is the bounded linear map $\dot{l}_\eta$ from $\mathcal{C} \to L_2^0(P_{\theta,\eta})$ with adjoint operator $\dot{l}_\eta^* : L_2^0(P_{\theta,\eta}) \to \mathcal{C}^*$. Note that $A^* : L_2^0(P_{\theta,\eta}) \to (\mathbb{R}^k \times \mathcal{C})^*$, the adjoint operator of A, is given by $A^*u\,(h,\beta) = \langle u\,,\,A(h,\beta)\rangle = h^T\langle \dot{l}_\theta\,,\,u\rangle_{P_{\theta,\eta}} + \dot{l}_\eta^*(u)(\beta)$.

Let the functional $\chi : \Theta \to \mathbb{R}^m$ with $m \le k$ be differentiable with respect to θ, and let $\chi'(\theta)_{m\times k}$ be its derivative. Consider the score operator for θ, $\dot{l}_\theta = (\dot{l}_{\theta 1}, \ldots, \dot{l}_{\theta k})^T$. Now for each $j = 1, \ldots, k$ write $\dot{l}_{\theta j} = \dot{l}^*_{\theta j} + \dot{l}^{**}_{\theta j}$ where $\dot{l}^{**}_{\theta j}$ is the orthogonal projection of $\dot{l}_{\theta j}$ into the closure of the range of $\dot{l}_\eta$. Thus $\dot{l}^*_{\theta j}$ is the orthogonal projection of $\dot{l}_{\theta j}$ into $\overline{R(\dot{l}_\eta)}^{\perp}$. We call $\dot{l}^*_\theta = (\dot{l}^*_{\theta 1}, \ldots, \dot{l}^*_{\theta k})^T$ the efficient score function for θ, and the efficient information for θ is then given by

$$I_{\theta,\eta} = \int \dot{l}^*_\theta \dot{l}^{*^T}_\theta\, dP_{\theta,\eta}\,.$$

Consider now the map $\kappa : \mathcal{P} \to \mathbb{R}^m$ given by $\kappa(P_{\theta,\eta}) = \chi(\theta)$. Also consider $\chi(\theta)$ as a function of the full parameter; thus $\chi(\theta) = \mu(\theta, \eta)$. It is now easy to see that

$$\frac{\partial}{\partial t}\left(\kappa(P_{\theta+th,\eta_t})\right)\Big|_{t=0}=\frac{\partial}{\partial t}\left(\mu(\theta+th,\eta_t)\right)\Big|_{t=0}$$
$$=\lim_{t\to 0}\frac{\chi(\theta+th)-\chi(\theta)}{t}=\chi^{'}(\theta)\,h=\dot{\mu}\,(h,\beta)\,,$$

$\dot{\mu}$ being a bounded linear map from $\mathbb{R}^k\times\mathcal{C}\to\mathbb{R}^m$. As before, a necessary and sufficient condition for pathwise norm differentiability of κ is the existence of $\dot{g}_{i0}$'s in $\overline{\mathcal{R}(A)}$ for $i=1,\ldots,m$ such that $\dot{\mu}^*(e_i)=A^*\dot{g}_{i0}$ for each i. Using arguments similar to Van der Vaart (1991) we can easily show that the necessary and sufficient condition translates to $\mathcal{N}(I_{\theta,\eta})\subset\mathcal{N}(\chi^{'}(\theta))$, which is the case if $I_{\theta,\eta}$ is invertible. In what follows, we assume that this is the case. We also assume that $\chi^{'}(\theta)$ is of full row rank (m). The efficient influence function $g_0=(\dot{g}_{10},\ldots,\dot{g}_{m0})$ (as before identified with the values at the basis vectors) is then given by $g_0=\chi^{'}(\theta)I_{\theta,\eta}^{-1}\dot{l}_\theta^*$, and the dispersion matrix for g_0, which acts as the information bound for the estimation of χ, is given by $J=\chi^{'}(\theta)I_{\theta,\eta}^{-1}\chi^{'}(\theta)^T$ and is invertible under our assumptions.

Denote the parameter (θ,η) by ψ. Consider the problem of testing the null hypothesis $\chi=\chi_0$ against $\chi\neq\chi_0$. Let $(\theta_0,\eta_0)\in\mathcal{H}_0$ be the true value of the parameter. Assume that all conditions of Theorem 2.1 hold and that (2.1) is satisfied with $I_0^{-1}\tilde{l}=g_0$. Here χ plays the role of θ in Theorem 2.1 while (θ,η) plays the role of ψ; thus we have $\sqrt{n}\,(\hat{\chi}-\chi_0)=\sqrt{n}I_0^{-1}\mathbb{P}_n\tilde{l}+o_p(1)$ under ψ_0 with $I_0=\left(\chi^{'}(\theta)I_{\theta,\eta}^{-1}\chi^{'}(\theta)^T\right)^{-1}=J^{-1}$ and $\tilde{l}=I_0g_0$. Now, consider local alternatives of the form

$$\psi_n=\psi_0+\frac{h}{\sqrt{n}}+o\left(\frac{1}{\sqrt{n}}\right)=(\theta_0+h_1/\sqrt{n},\eta_0+h_2/\sqrt{n}+o(1/\sqrt{n}))\,.$$

Using Theorem 2.2 we readily deduce that under ψ_n, $\mathrm{lrt}_n(\chi_0)\to_d\chi_m^2\,(c^TIc)$ with $c=I_0^{-1}\langle Ah\,,\,\tilde{l}\rangle_{P_{\psi_0}}$ and $h=(h_1,h_2)$. Since χ, considered as a functional defined on the space of probabilities (the map κ), is pathwise norm differentiable at ψ_0, by the assumption of invertibility of I_{θ_0,η_0} we have that

$$c=\frac{\partial}{\partial t}\left(\kappa(P_{\theta_0+th_1,\eta_0+th_2+o(t)})\right)\Big|_{t=0}=\chi^{'}(\theta_0)\,h_1\,.$$

Thus the final form for the non-centrality parameter Δ is

$$\Delta=h_1^T\chi^{'}(\theta_0)^T\left(\chi^{'}(\theta_0)I_{\theta_0,\eta_0}^{-1}\chi^{'}(\theta_0)^T\right)^{-1}\chi^{'}(\theta_0)h_1\,.$$

When $\chi(\theta)=\theta$ (which happens in many cases), $\chi^{'}(\theta_0)$ is the identity matrix and the expression for the non-centrality parameter reduces to $\Delta=h_1^TI_{\theta_0,\eta_0}h_1$;

here I_{θ_0,η_0} is the efficient information for estimating θ at parameter value ψ_0. This expression is exactly what one gets in the regular parametric case, where η is finite dimensional, since the discussion above includes the special case when η, the nuisance parameter varies in a finite dimensional space.

Application to the Cox Proportional Hazards Models. The Cox Proportional Hazards Model is probably the most widely used of all semiparametric models. The distribution of the likelihood ratio statistic for testing $H_0 : \theta = \theta_0$ in the Cox Proportional Hazards setting has been studied under various censoring schemes. Two important schemes are the following: (i) Right Censoring – here we observe $X = (T \wedge C, 1\{T \le C\}, Z)$ where, given (the covariate) Z, the variables T (survival time) and C (observation time) are independent and T follows the Cox Model. (ii) Interval Censoring – here we observe $X = (1\{T \le C\}, Z)$; thus, we have less information on the survival time distribution compared to the right censoring scenario. For both right censoring and interval censoring, θ is estimable at $\sqrt{n}$ rate by its MLE which is asymptotically normal and efficient, and the likelihood ratio test for testing $H_0 : \theta = \theta_0$ converges under the null hypothesis to a χ^2 distribution. With right-censoring Λ, the baseline hazard, can be estimated at $\sqrt{n}$ rate (by its MLE), but with interval censoring the rate slows to $n^{1/3}$. For more details on likelihood based estimation and inference on the Cox Model with right censored data see, for example, Van der Vaart (1998), Banerjee (2000) and Murphy and Van der Vaart (2000). Maximum likelihood estimation in the Cox model with interval censored data was first considered in Huang (1996), where the MLE of θ was shown to asymptotically normal and efficient. The likelihood ratio statistic for testing $H_0 : \theta = \theta_0$ was shown to be asymptotically χ^2_d (d being the dimension of θ) under the null hypothesis by Murphy and Van der Vaart (1997, 2000), under appropriate regularity conditions. Natural local alternatives to consider in the Cox PH setting with right-censored data are of the form $(\theta_0 + h_1/\sqrt{n}\,, \Lambda_0 + (1/\sqrt{n}) \int_0^{\cdot} h_2 \, d\Lambda_0)$, where h_2 is a bounded function in $L_2(\Lambda_0)$. Using the ideas of this paper, we can show (under appropriate assumptions) that $\mathrm{lrt}_n(\theta_0)$, the likelihood ratio statistic for testing $\theta = \theta_0$, converges in distribution, under the sequence of local alternatives above, to $\chi^2_d(h_1^T I_0 \, h_1)$, where I_0 is the efficient information for θ in the right censored problem at parameter value (θ_0, Λ_0). For the details, see Banerjee (2000). In the Cox PH setting with interval–censored data, the natural local alternatives take the form $(\theta_0 + h_1/\sqrt{n}, \Lambda_0 + h_2/\sqrt{n})$ where h_2 is a non-negative non-decreasing continuous function. Once again, it can be shown that the power of the likelihood ratio statistic for testing $H_0 : \theta = \theta_0$ converges, under the above sequence of local alternatives, to $\chi^2_d(h_1^T \, \tilde{I}_0 \, h_1)$; here $\tilde{I}_0$ is the efficient information for θ in the interval–censoring set-up. For an explicit representation of $\tilde{I}_0$, see Huang (1996).

For more concrete applications of the results in this paper, we refer the reader to Chapter 2 of Banerjee (2000).

Acknowledgements

Thanks to Susan Murphy and Jon Wellner for helpful comments and discussions. This research was supported partially by National Science Foundation grants DMS-0071818 and DMS-0306235.

References

Banerjee, M. (2000). Likelihood ratio inference in regular and nonregular problems. Ph.D. dissertation, University of Washington.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models.* Springer Verlag, New York.

Engle, R., Granger, C., Rice, J. and Weiss, A. (1986). Semiparametric estimation of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* **81**, 310-320.

Green, P. (1984). Iteratively reweighted least squares for some robust and resistant alternatives. *J. Roy. Statist. Soc. Ser. B* **46**, 149-192.

Huang, J. (1996). Efficient estimation for the Cox model with interval censoring. *Ann. Statist.* **24**, 540-568.

Murphy, S. and Van der Vaart, A. (1997). Semiparametric likelihod ratio inference. *Ann. Statist.* **25**, 1471-1509.

Murphy, S. and Van der Vaart, A. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95**, 449-465.

Murphy, S. and Van der Vaart, A. (2001). Semiparametric mixtures in case control studies. *J. Multivariate Anal.* **79**, 1-32.

Roeder, K., Carroll, R. J. and Lindsay, B. G. (1996). A semiparametric mixture approach to maximum likelihood estimation in frailty models. *Scand. J. Statist.* **19**, 25-44.

Van der Vaart, A. (1991). On differentiable functionals. *Ann. Statist.* **19** , 178-201.

Van der Vaart, A. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer, New York.

Van der Vaart, A. (1998). *Asymptotic Statistics.* Cambridge University Press, Cambridge.

University of Michigan, Department of Statistics, 439, West Hall, 1085 South University, Ann Arbor, MI, 48109, U.S.A.

E-mail: moulib@umich.edu

(Received May 2003; accepted May 2004)